Unfolding: Introduction

Louis Lyons
Imperial College, London
The problem

Given a one-dimensional histogram for a particular variable, obtained in a detector with known experimental resolution, can we estimate the distribution that would have been obtained if the detector did not introduce any smearing? And we need to provide a covariance matrix for our unsmeared distribution.

(Some obvious generalisations of this)
Programme

9.10 Victor Panaretos "A statistician's view"
10.00 Volker Blobel "Unfolding methods for particle physics"
10.40 Coffee, and Poster Session
11.10 Guenter Zech "Regularization and error assignment in unfolding"
11.40 Vato Kartvelishvili "Unfolding with SVD"
12.10 Katharina Bierwagen 'Bayesian Unfolding'
12.25 Comparison on HEP methods
12.45 Lunch

2.00 Bogdan Malaescu 'Iterative dynamically stabilized method of data unfolding'
2.25 Kerstin Tackmann 'SVD-based unfolding: implementation and experience'
2.45 Michael Schmelling 'Regularization by control of resolution function'
3.05 Hans Dembinski 'ARU - towards optimal unfolding of data distributions'
3.30 Tim Adye 'Unfolding algorithms and tests using RooUnfold'
3.55 Coffee
4.25 Matthias Weber 'CMS unfolding'
4.55 Georgios Choudalakis 'ATLAS unfolding'
5.25 Jan Fiete Grosse-Oetringhaus 'ALICE unfolding'
5.45 Summary (Victor Panaretos) + Discussion
Non-expert comments on Unfolding

• Why unfold?
• Correction factors
• Choice of bin-size
• How good is your method?
• Error estimates
Why Unfold?

If possible, **don’t Unfold data**, but smear theory

When not possible?

a) Compare data with data from different experiment

b) Tune MC by fitting QCD parameters to data

c) Future theories

Also provides more useful result (but complicated correlated errors)
Why Unfold?

If possible, **don’t Unfold data**, but smear theory

When not possible?

a) Compare data with data from different experiment
b) Tune MC by fitting QCD parameters to data
   
   **c) Future theories**  
   Provide smearing matrix, so that future theorists can smear their theories

Also provides nicer picture to look at (but complicated correlated errors)
Matrix method

\[ d_i = \Sigma M_{ij} t_j \]

Assume \( M_{ij} \) known with small statistical error from MC \{Estimate effect of \( M_{ij} \) bias from incorrect resolution\}

Max likelihood solution can have large bin-to-bin oscillations. Effect not serious for wide enough bins
Bin-by-bin correction factors

Use MC to find how exptl resolution makes:

‘Truth’ → Observed data

Beware MC statistical and systematic errors

\[ t_i = C_i \times d_i \]  
(Contrast \[ t_i = \sum M^{-1}_{ij} \times d_j \] for matrix method)

Problems:

1) \( C_i = 0.1, \, d_i = 100 \pm 10 \) → \( t_i = 10 \pm 1 \) ??
   i.e. Error too small. But this error estimate is incorrect

2) \( C_i \) depends on assumed distribution of \( t \), which we are trying to find.
   (For small bin-size, matrix method is less sensitive to distribution of \( t \))

3) No bin-to-bin estimates of correlations

4) Sum of estimated truth ≠ Sum of observed data (Matrix method O.K.)

CONCLUSION: Do not use  (Cf Cousins)
## Problems with Correction Factors

<table>
<thead>
<tr>
<th>Truth (not known)</th>
<th>Mean observed</th>
<th>Bin 1</th>
<th>Bin 2</th>
<th>Smearing matrix</th>
<th>True</th>
</tr>
</thead>
<tbody>
<tr>
<td>800</td>
<td>200</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>760</td>
<td>240</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Smearing matrix

<table>
<thead>
<tr>
<th>Observed</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>0.1</td>
<td>0.8</td>
</tr>
</tbody>
</table>

### Assumptions and Calculations

- **As \( t_i \)**: Assumed # in bin \( i \)
- **Re \( d_i \)**: Resulting expected # in bin \( i \)
- **CF \( i \)**: Correction factor for bin \( i \) = (As \( t \)) / (Re \( d \))
- **Est \( t_i \)**: Estimated true # in bin \( i \)
- **Est \( t_1 \)**: \( 760 \times CF_1 \) (truth = 800)
- **Est \( t_2 \)**: \( 240 \times CF_2 \) (truth = 200)

<table>
<thead>
<tr>
<th>Bin 1</th>
<th>Bin 2</th>
<th>As ( t_1 )</th>
<th>As ( t_2 )</th>
<th>Re ( d_1 )</th>
<th>Re ( d_2 )</th>
<th>CF ( 1 )</th>
<th>CF ( 2 )</th>
<th>Est ( t_1 )</th>
<th>Est ( t_2 )</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>0</td>
<td>900</td>
<td>100</td>
<td>1.11</td>
<td>0</td>
<td>844</td>
<td>0</td>
<td>844</td>
<td></td>
<td></td>
</tr>
<tr>
<td>800</td>
<td>200</td>
<td>760</td>
<td>240</td>
<td>1.05</td>
<td>0.83</td>
<td>800</td>
<td>200</td>
<td>1000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>760</td>
<td>240</td>
<td>732</td>
<td>268</td>
<td>1.04</td>
<td>0.90</td>
<td>789</td>
<td>215</td>
<td>1004</td>
<td></td>
<td></td>
</tr>
<tr>
<td>667</td>
<td>333</td>
<td>667</td>
<td>333</td>
<td>1.00</td>
<td>1.00</td>
<td>760</td>
<td>240</td>
<td>1000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>500</td>
<td>550</td>
<td>450</td>
<td>0.91</td>
<td>1.11</td>
<td>691</td>
<td>267</td>
<td>958</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>800</td>
<td>340</td>
<td>660</td>
<td>0.59</td>
<td>1.21</td>
<td>447</td>
<td>291</td>
<td>738</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1000</td>
<td>200</td>
<td>800</td>
<td>0</td>
<td>1.25</td>
<td>0</td>
<td>300</td>
<td>300</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( \text{Est } t_1 = 760 \times CF_1 \) (truth = 800)

\( \text{Est } t_2 = 240 \times CF_2 \) (truth = 200)
Bin size?

Not too small

\[ M_{ij} \] has large off-diagonal elements

Not too large

Lose sensitivity

\[ M_{ij} \] begins to depend on true distribution

Bin width for unfolded distribution also depends on that for data, and on exptl resolution.

Recommendation for optimum?
Regularisation

Damps out oscillations at price of (small?) bias

Recipe for optimal regularisation? (Depends on....)

How to judge which method is ‘best’? Challenge?

Bob Cousins’ “bottom line test”:

Compare 2 theories with data via $\chi^2$ ($\chi^2_1$ and $\chi^2_2$)

a) by smearing theories (well-defined)

b) by unsmearing data (use your favourite method)

Do $\Delta \chi^2 = \chi^2_1 - \chi^2_2$ for a) and for b) agree?
Can unfolded dist have $\sigma < \sqrt{n}$?

Regularisation can produce this.

Cf Straight line fitting to data $\rightarrow$ fitted uncertainties smaller than measured ones

Estimate errors by MC or by bootstrap replications of data
Looking forward to all the talks, and to hearing what Statisticians (especially Victor Panaretos) can tell us about the subject.
Looking forward to all the talks, and to hearing what Statisticians (especially Victor Panaretos) can tell us about the subject. Maybe even consensus on ‘best approach’?